Fermilab U.S. DEPARTMENT OF ENERGY | Office of Science

Introduction to HEP Computing in Fermilab Scientific Computing Division

Pengfei Ding
Scientific Computing Division, Fermilab

New Perspective 2016

14 June 2016

Instead of giving an overview of Scientific Computing in Fermilab, I am going to talk about:

How does computing take part in a HEP graduate student's or postdoc's daily work?

HEP students'/postdocs' physics life

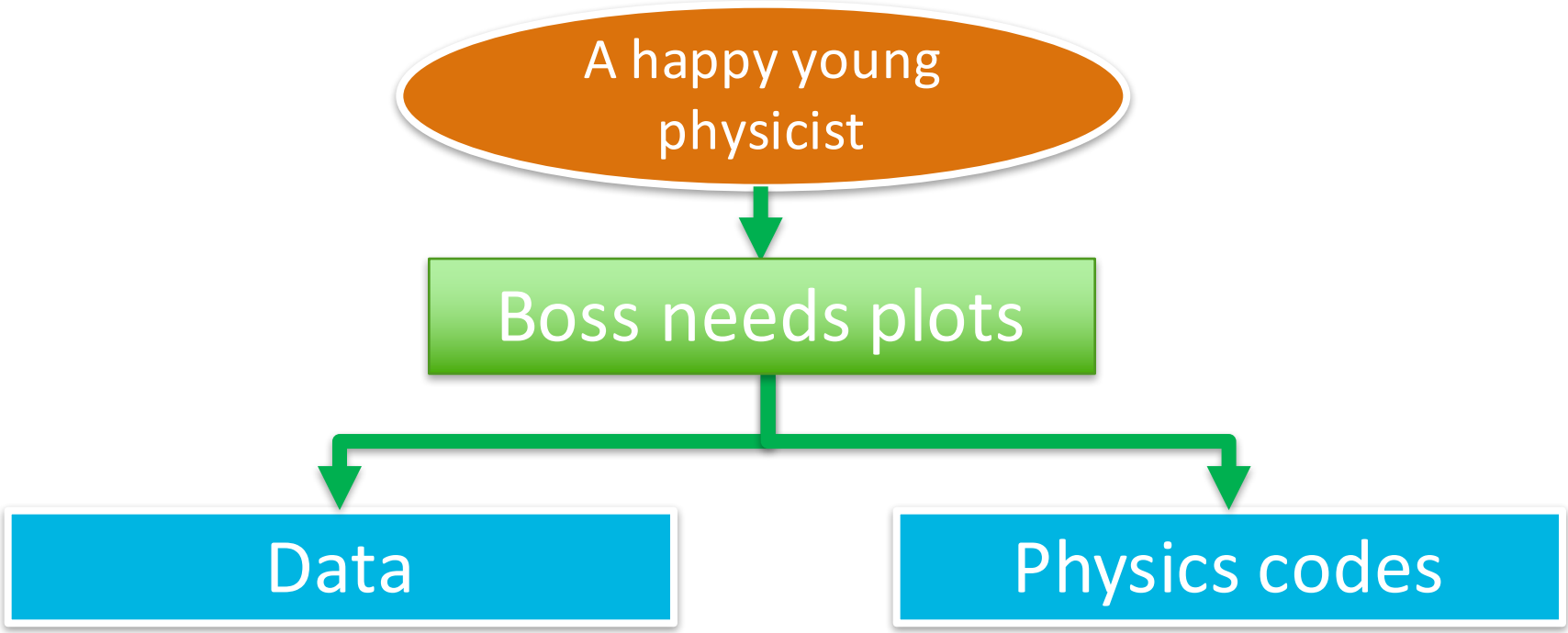
A happy young physicist

HEP students'/postdocs' physics life

A happy young physicist

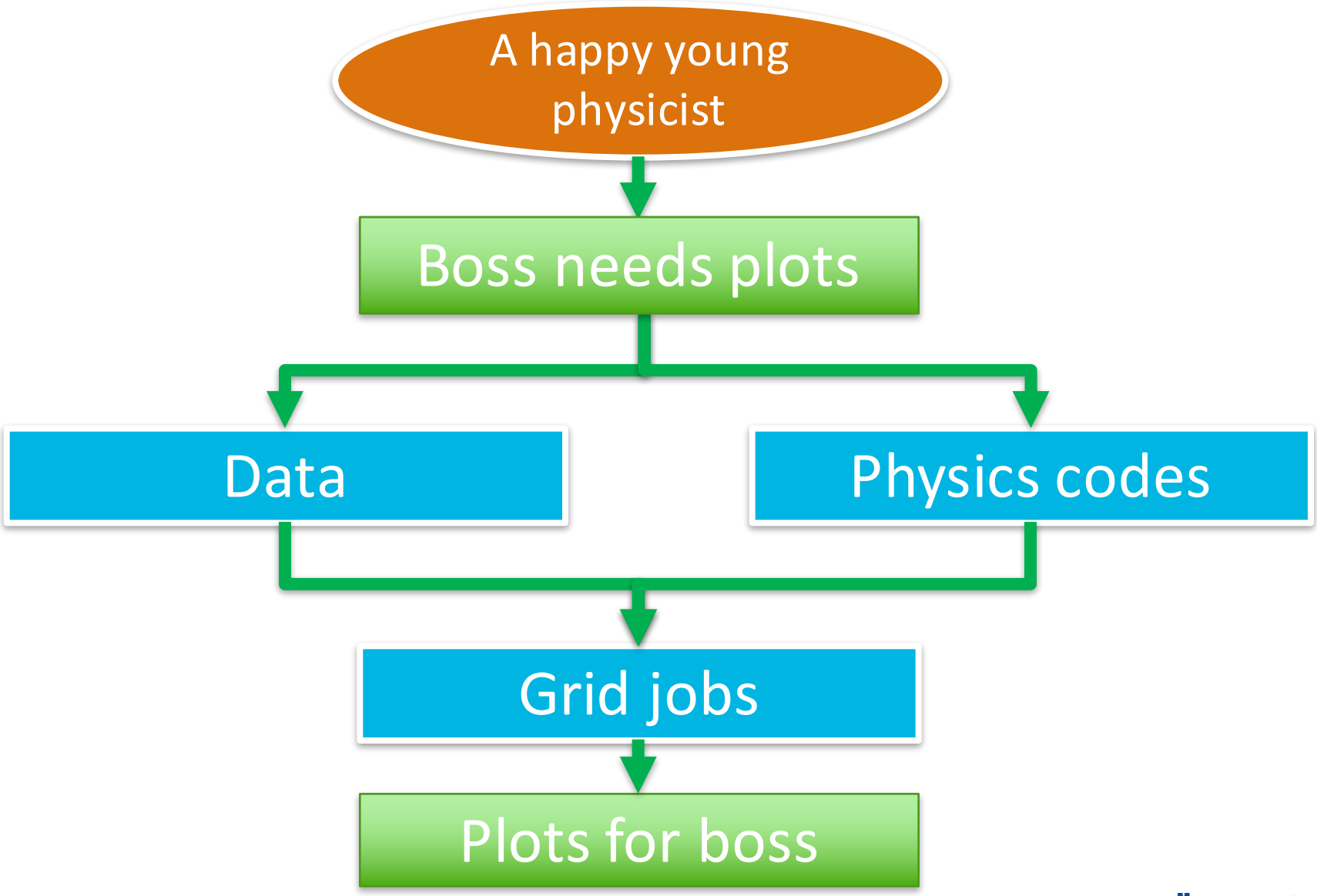
Boss wants plots…
What should I do?

HEP students/postdocs workflow



“But processing the data on my laptop takes ages and there is so much data.”

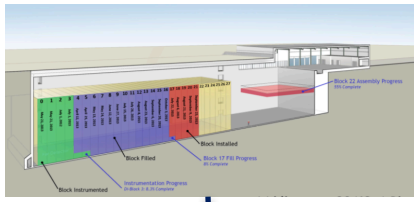
HEP students/postdocs workflow



Many things happened behind the scene

- How is data acquired from detector?
- How is data transferred, stored and accessed?
- How does my physics code integrate with the framework?
- Where do the submitted jobs run?

Overview of HEP Computing workflows (I)



Data Acquisition from Detector

Online processing

Buffering and triggering

Data Transfer from detector to central storage site

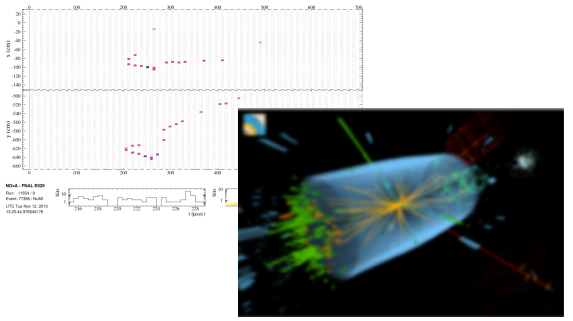
Data Catalog and archival

"Tier-1" Facility for storage, reprocess, local access

Processing Infrastructure: data/MC

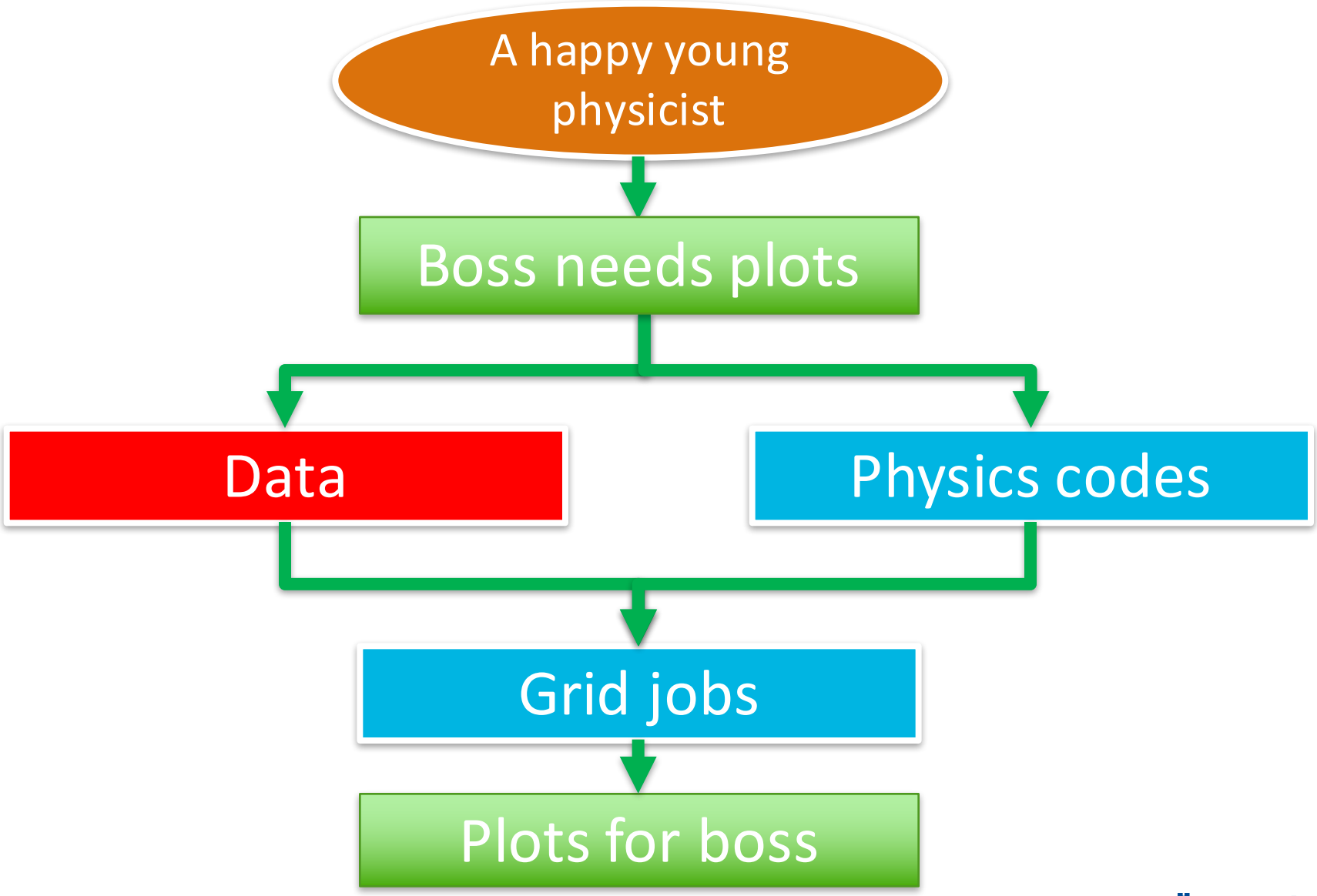
Analysis Facility FermiGrid, OSG

Physics results



SCD

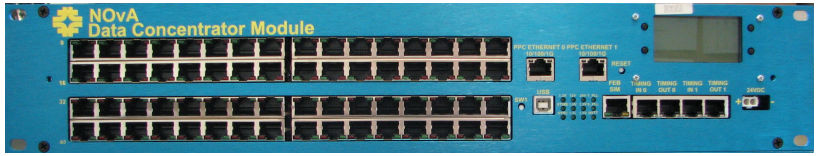
HEP students/postdocs workflow



“Where does data come from?”

Data (I) - Data Acquisition (DAQ)

- Readout electronics:
  - SCD electrical engineers in real-time application group, specialized in FPGAs, ASICs, embedded systems, designs readout electronics and develops firmware.

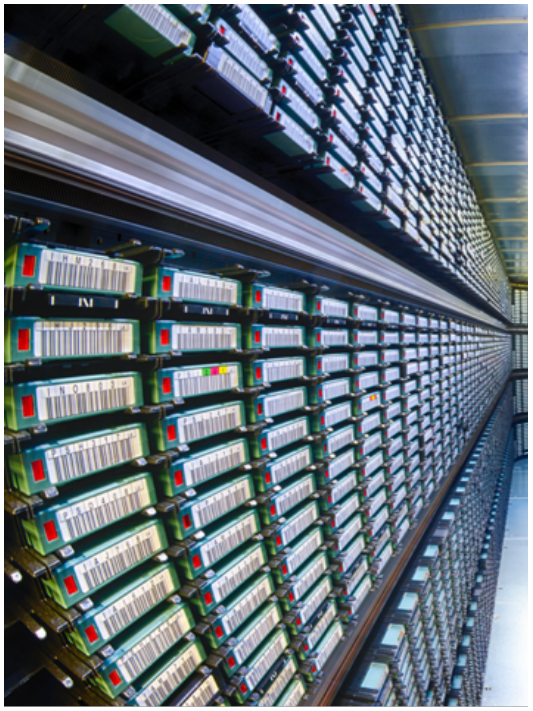


- Online software for DAQ:
  - ***Artdaq***:
    - a real-time software system for data acquisition
    - Integration of offline code to online system
    - Used by Mu2e, SBND, protoDUNE, DarkSide …

“How to store so much data?”

Data (II) – Data Storage – Tape ("Enstore")

Tape Storage



Cost effective way to store huge amount of data for years

Capacity: *(equivalent to 2 million laptops with 512 GB hard drive)*

- 7x 10,000 slot libraries
  - With 5.4 TB T10000C ~ 375 PB
  - With 8.5 TB T10000D ~ 595 PB
- Allocation:
  - General purpose: 4 libraries
  - CMS: 3 libraries

- Current usage
  - General purpose ~ 22 PB
  - CMS ~ 43 PB
  - Legacy CDF, DZero ~ 30 PB
  - (includes migration duplicates)

“Tape is too slow. Boss cannot wait. And I can not graduate within a decade with data only in tape. What shall I do?”

Data (III) – Data Storage – Cache disks (“dCache”)

Cache disk (dCache):*(equivalent to ~60, 000 laptops with 512 GB hard drive)*

- General purpose ~ 8.5 PB
- CMS ~ 22 PB
- Legacy CDF ~ 1.5 PB

Project/user disk (NAS):

- General purpose ~ 2 PB
- CMS (EOS) ~ 4 PB
- Legacy CDF, DZero ~ 1 PB

HPC disk (Lustre):

- LQCD ~ 1 PB

Disk Systems

“How do I find data in the huge storage facilities?”

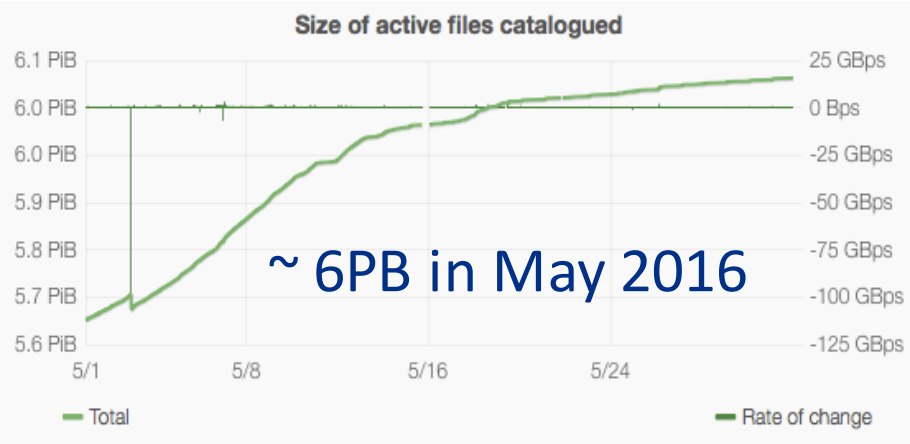
-File names? That works for small amount of files.

-What if I don’t know the file names? -What if I want to find all the data taken between a specific time period?

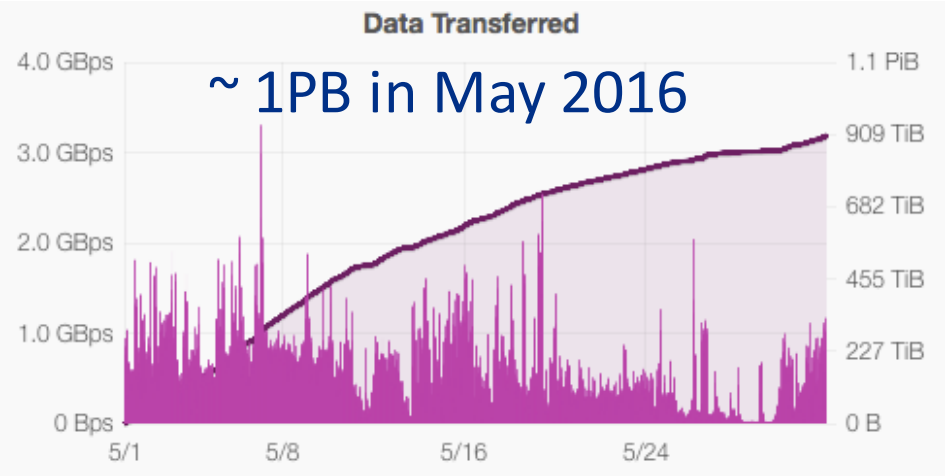
Data (IV) – Data Catalog - SAM

- **SAM**: handles the file metadata and bookkeeping.
  - bookkeeping information is kept in a database, which can be accessed by a series of commands.
  - to find data of interest, the user doesn't need to know where the files are, or what the individual files are.

Files catalogued by SAM for NOvA

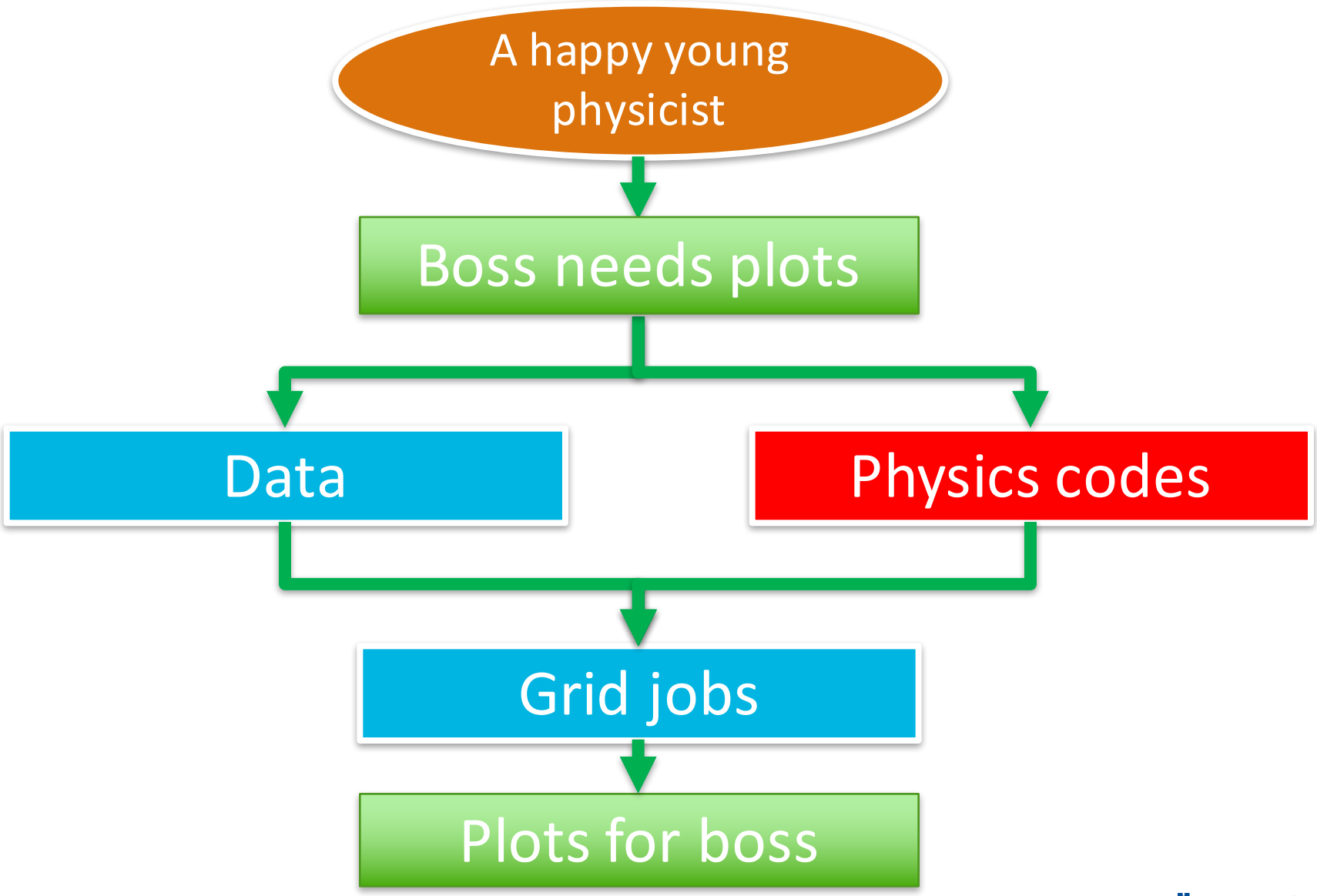


Data transferred by File Transfer Service



Fermilab

HEP students/postdocs workflow



Software framework (I)

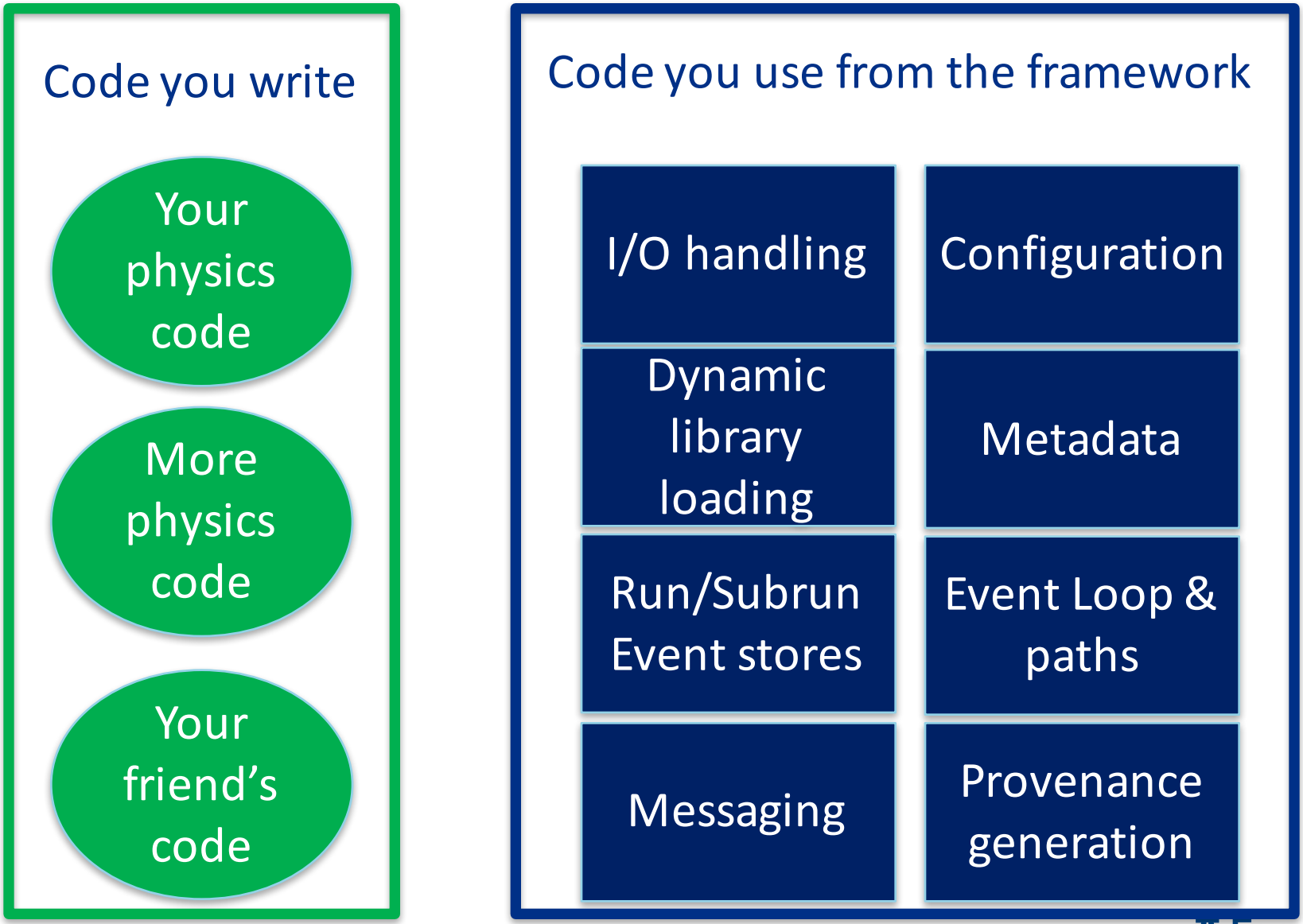
1) Loop over events and reconstructed objects;

2) Apply physics cuts;

3) Fill histograms.

```
// Slices
art::Handle< std::vector< rb::Cluster > > slices;
e.getByLabel(fSlicerLabel, slices);

double nNoiseSlices  = 0.;
double nNoiseHits    = 0.;
double nHitsInSlices = 0.;

const unsigned int nSlices = slices->size();

for(unsigned int sliceIdx = 0; sliceIdx < nSlices; ++sliceIdx){
    art::Ptr<rb::Cluster> aSlice(slices, sliceIdx);

    int    min_cell_sep  = 999;
    double min_plane_sep = 999;
    double min_sep       = 999;

    if(aSlice->IsNoise()){
        ++nNoiseSlices;
        nNoiseHits += aSlice->NCell();

        hNoiseSlice_hits_vs_mb->Fill(nEvent, aSlice->NCell());
        hNoiseMeanADC_vs_Slice_hits->Fill((aSlice->TotalADC() / (double)aSlice->NCell()), aSlice->NCell());
        hNoiseSlicesNHits->Fill(aSlice->NCell());
        hNoiseSlicesSumADC->Fill(aSlice->TotalADC());
        hNoiseSlicesMeanADC->Fill((aSlice->TotalADC() / (double)aSlice->NCell()));
        //FillPlaneSeperations(aSlice, min_cell_sep, min_plane_sep, min_sep);
        hNoiseSlicesMinInPlaneSeperation->Fill(min_cell_sep);
        hNoiseSlicesMinCrossPlaneSeperation->Fill(min_plane_sep);
        hNoiseSlicesMinSeperation->Fill(min_sep);
    } else {
        nHitsInSlices += aSlice->NCell();

        hSlice_hits_vs_mb->Fill(nEvent, aSlice->NCell());
```

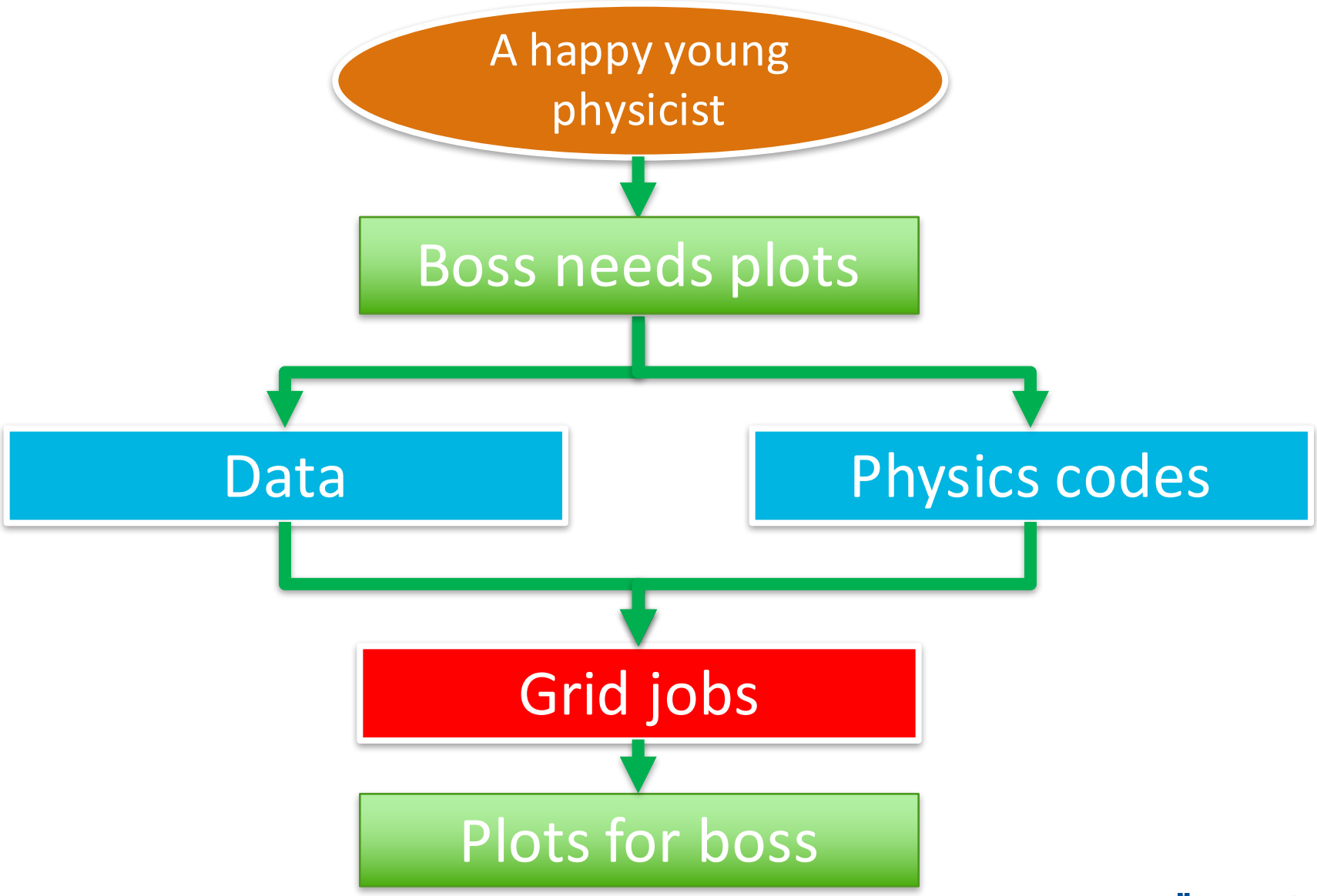
Software framework (II)



Software framework (III)

- ***art,*** a software framework
  - let physicists write physics code only
  - development and support is shared among multiple experiments
  - used by NOvA, Mu2e, DUNE, LArIAT, MicroBooNE, g-2, SBND, DarkSide and NEXT
- ***LArSoft***:
  - Liquid Argon TPC experiments use similar simulation and reconstructions methods
  - jointly developed by all Liquid Argon TPC experiments
  - a toolkit built on top of ***art***
  - provides additional toolkits for simulation and reconstruction designed specifically for Liquid Argon TPC experiments

HEP students/postdocs workflow



“Where do my grid jobs run?”

Computing Facilities (I) – FermiGrid and OSG

FermiGrid: (mostly used by users currently)
It is for Fermilab experiments;

Open Science Grid - OSG:
It is shared and contributed by many institutions and other fields than HEP.

FermiGrid "worker node" core counts:
General purpose:

- 16,608 cores (GP Grid)

Also manage:

- 17,872 cores (CMS Tier-1)
- 4,984 cores (CMS LPC)
- 28,240 cores (HPC / LQCD)
- 2,008 cores (DZero legacy)

FIFE

May 2016

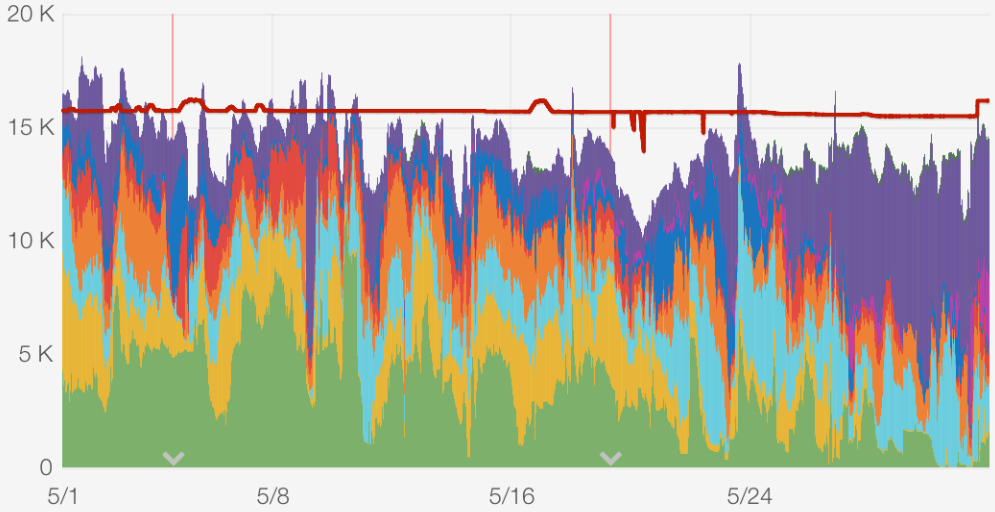
Average Number of Jobs Running Concurrently

14040

Total Jobs Run

3913645

Running Jobs by Experiment (includes Onsite, OSG & Cloud)



	min	max	avg
NOvA	0	11.72 K	3.65 K
Mu2e	0	6.00 K	1.94 K
MINERvA	0	7.09 K	2.17 K
MINOS	0	5.46 K	1.64 K
DUNE	0	5.12 K	947
MicroBooNE	0	5.51 K	985
DES	0	6.03 K	161
Other Experiments	0	10.29 K	2.54 K
Projects	0	39	6
Onsite Slots (GPGrid)	13.94 K	16.22 K	15.69 K

Percent Jobs Run Onsite

89.7%

Percent Jobs Run on HEP Cloud

0.0%

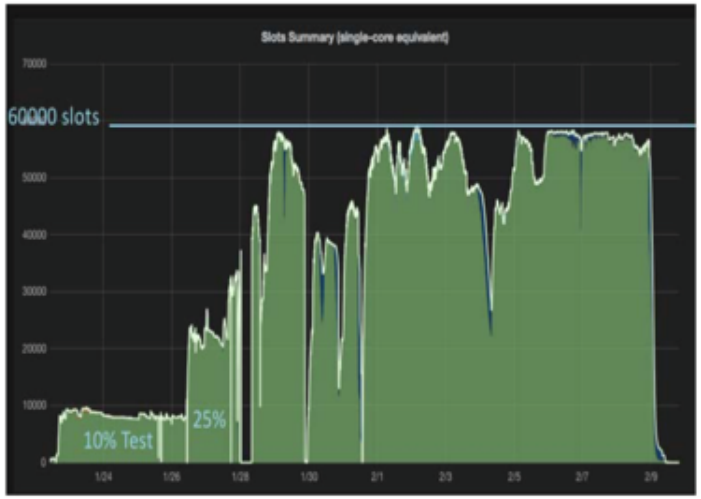
Percent Jobs Run on OSG

10.3%

“What if we want more computing resources?”

Computing Facilities (II) – HEP Cloud

- HEPCloud – A new facility paradigm:
  - Provides “elastic” deployment of resources
  - A single protal to a heterogeneous set of resources, e.g clouds, grids, HPC etc, both local and ”rental” (commercial, such as Amazon AWS, Microsoft cloud, Google cloud)



A piloting test with HEPCloud for CMS during Moriond rush showed the ability of having 56,000 computing cores steady. Huge boost to the worldwide CMS computing capability.

25% boost

Fermilab

“It seems none of the above is related to me. I don’t use any of these.”

- An accelerator/cosmology/theory student

Computing Facilities (III) - High Parallel Computing

- Lattice QCD
  - Three Clusters
    - ~25,000 CPU cores total
    - Two clusters include GPUs

- Accelerator Simulation and Cosmology
  - Two clusters
    - ~2,500 CPU cores total

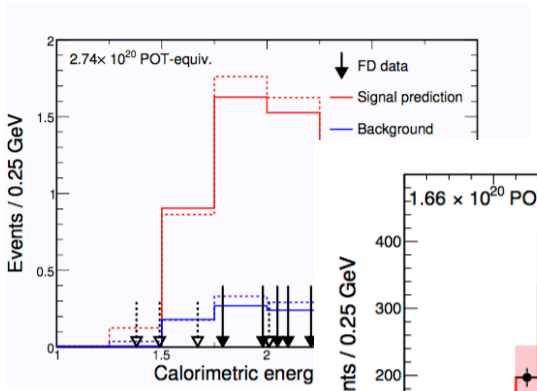
- Next-generation research and testing facilities
  - Two clusters of 72 traditional cores
  - GPU clusters
  - Intel Xeon Phi clusters

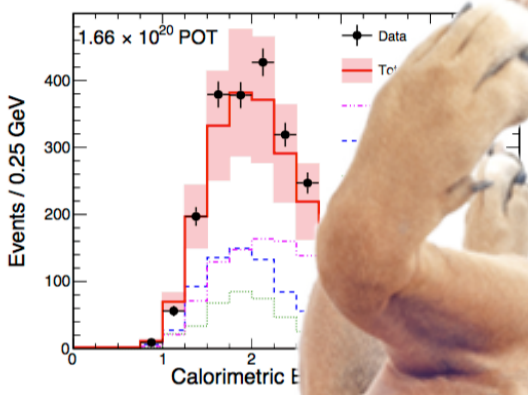
How do things work together?

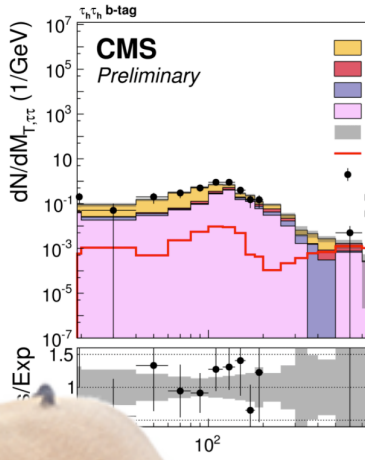
- How do things work together?
  - My local physics code
  - Huge amount of Data in tape and/or dCache
  - Thousands of CPUs provided by computing facilities: FermiGrid, OSG, HEPCloud

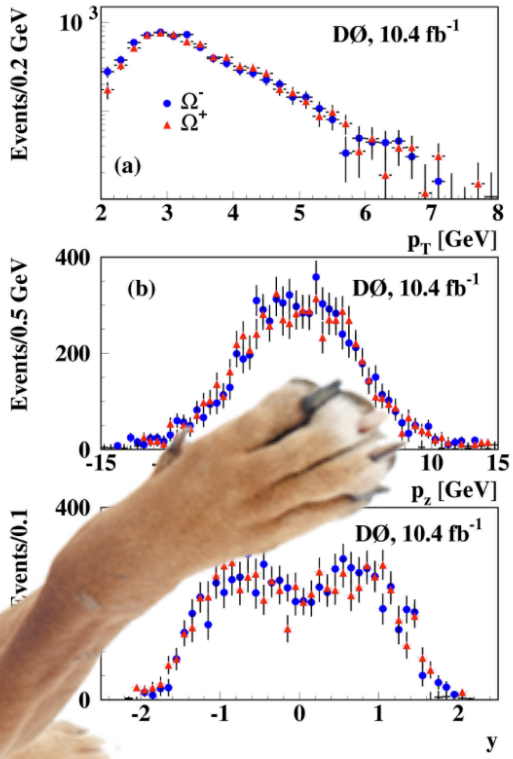
- **F**abr**I**c For **F**rontier **E**xperiments (**FIFE**): provides common computing services and interfaces
  - manages job submission to different computing facilities;
  - provides monitoring tools;
  - does data handling for jobs (deliver input files, catalog and transfer output files).

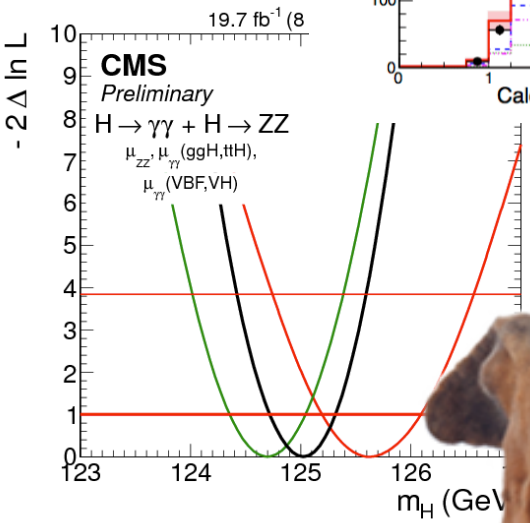
Things are working together!





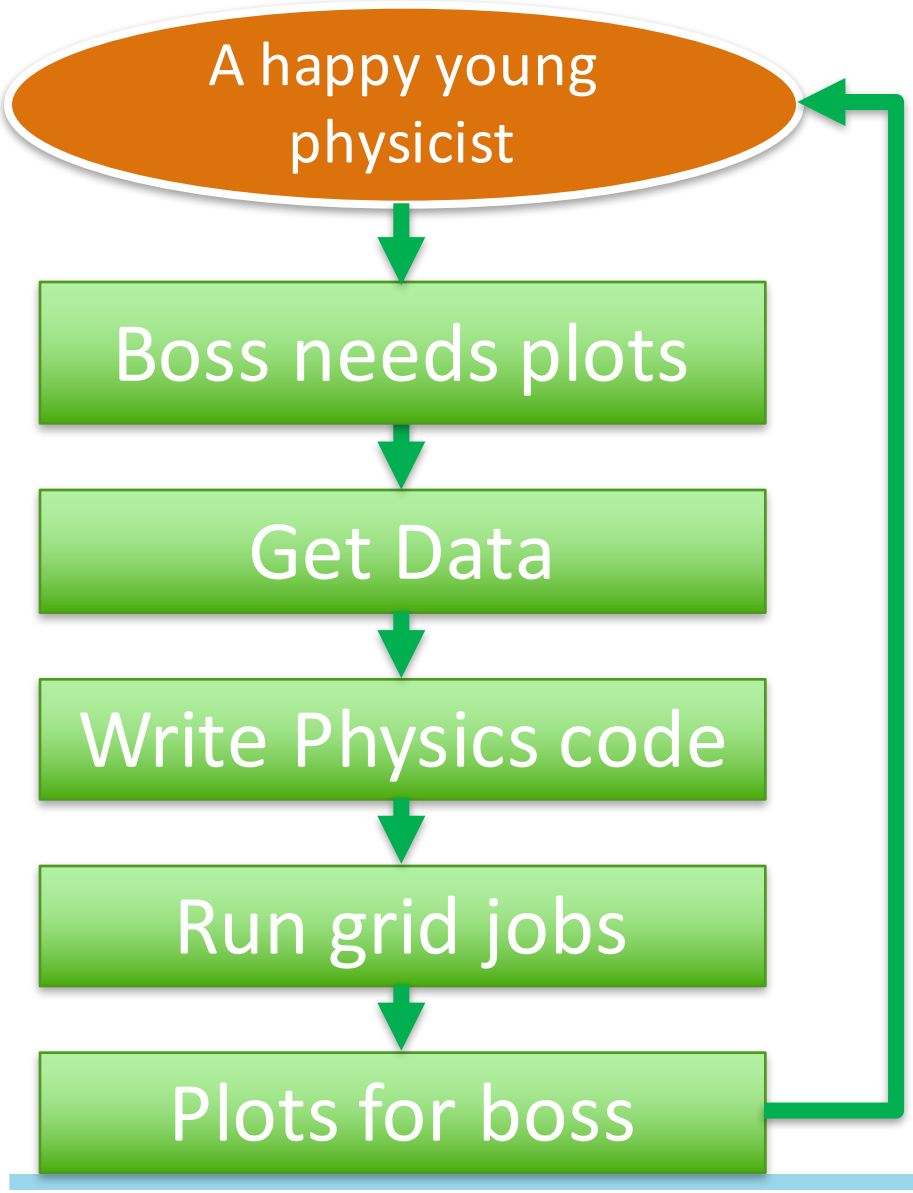






Fermilab

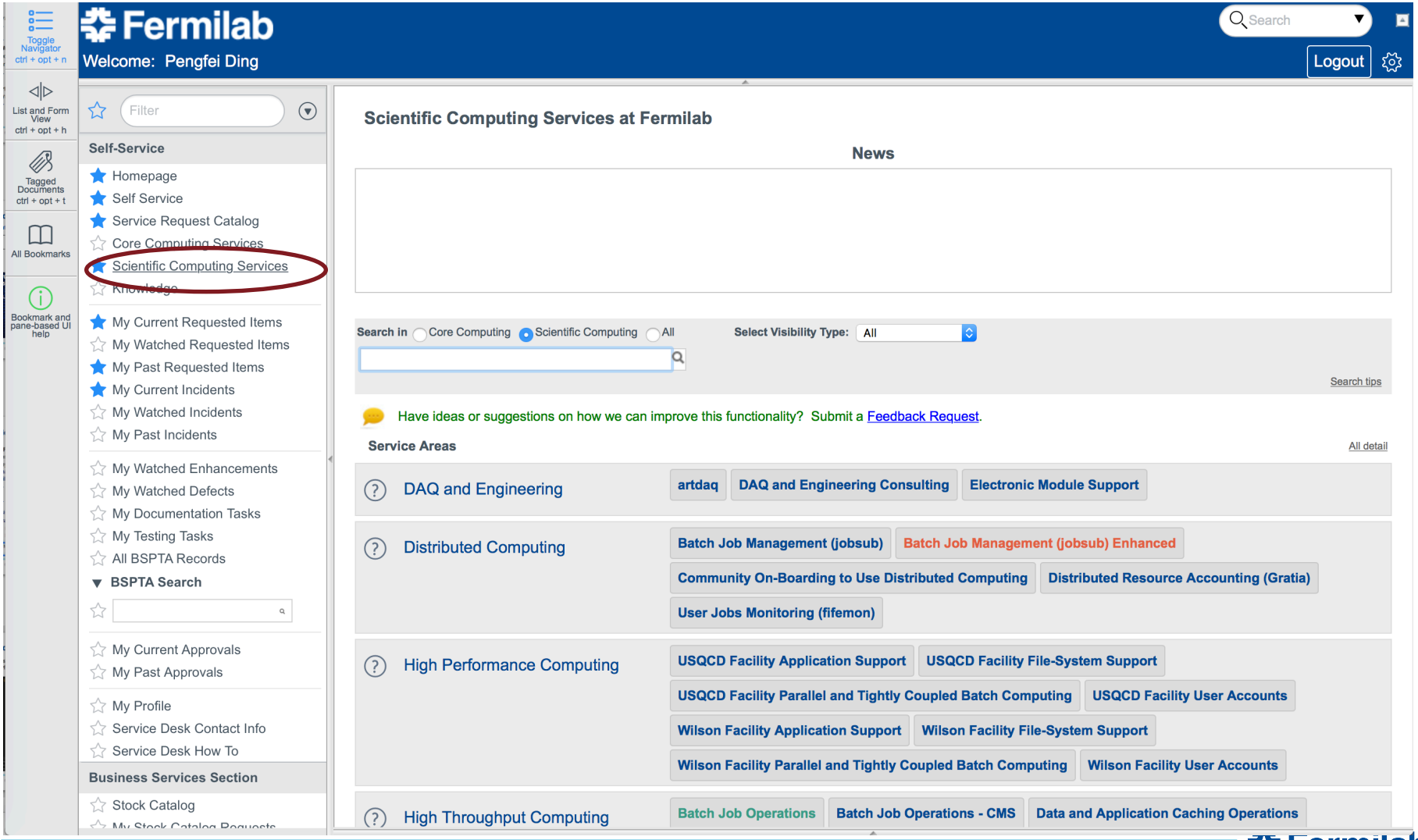
Summary



- With all the things happend behind the sceen, SCD wants to keep your physicst's life simple and free from many computing issues.

Submit a ticket when having computing issues

https://fermi.service-now.com/navpage.do



Backup

Data Handling (I)

Mass storage:

- **Enstore**: provides access to data on **tape** to/from a user's machine on-site, or over the wide area network through the dCache disk caching system
  - CDFEN for CDF RunII; D0EN for D0 RunII; STKEN for all other Fermilab users, including US-CMS Tier1
  - Consisting of 7x10,000 slot robotic tape libraries (1.5 PB/month, 50 TB/day, peak was 6PB/month)
- **dCache**: **disk** caching software. Fermilab dCache systems use raided disk in redundant configurations.
  - US CMS Tier1 dCache System; CDF dCache System; Public dCache System for all other Fermilab users
- **PNFS/Chinmera Namespace**: is used both by dCache and Enstore to distribute files names and other storage related metadata.

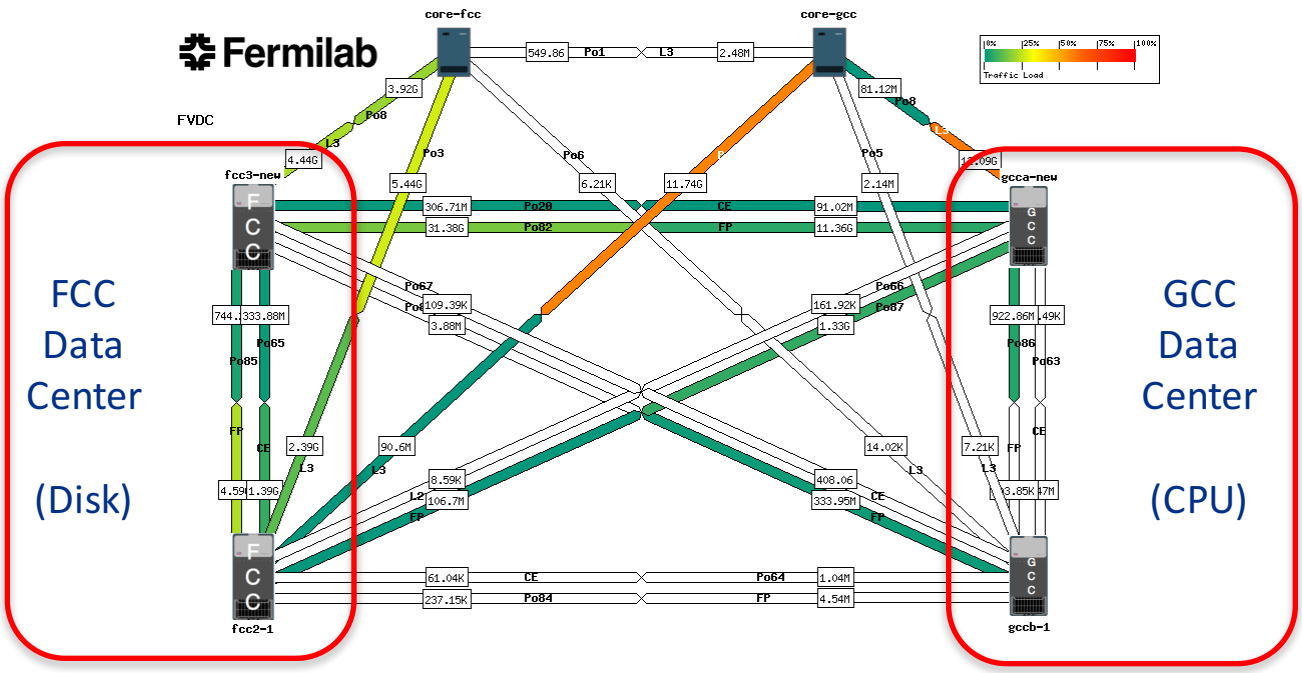
Computing Facilities (I) -

- Much of the computing in HEP relies on **H**igh **T**hroughput **C**omputing (**HTC**)
  - One task per processor on many processors
  - Trivial to perform in parallel
  - Facilities (farms of computing nodes): FermiGrid, OSG, HEPCloud,

- **H**igh **P**erformance **C**omputing (**HPC**)
  - One large task on many processors
  - Tightly-coupled communications • Advanced networking
  - Low latency and high bandwidth
  - Includes Linux clusters with specialized networking and supercomputers
  - Facilities: HPC clusters, next generation HPC testbeds

Facility resources - Network

Offsite network: 2x 100Gb, 3x 10Gb



Networks

Network support is provided by Core Computing Division, with effort and M&S funded by KA22 02 Detector Operations and KA21 02 Energy Facilities (CMS)

Fermilab